DOCUMENT RESUME

ED 205 595 TM 810 496

AUTHOR Stewart, E. Elizabeth
TITLE Methodological Issues Related to the Study of Context Effects in Multisection Tests.
PUB DATE 14 Apr 81
NOTE 12p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 14, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Collection: *Error of Measurement: *Evaluation Methods: Factor Analysis: Orthogonal Rotation: *Performance Factors: *Test Format: Testing: Test Reliability: Test Theory
IDENTIFIERS *Context Effect: Operational Questions: Standard Deviation

ABSTRACT

Context effects are defined as being influences on test performance associated with the content of successively presented test items or sections. Four types of context effects are identified: (1) direct context effects (practice effects) which occur when performance on items is affected by the examinee having been exposed to similar types of material earlier in the test: (2) indirect context effects occurring when the examinee has been exposed earlier in the test to a different type of material: (3) serial position effects which result from the amount, rather than the type, of material that has preceded a particular section: and (4) motivational effects, when the examinee believes that a particular section is non-operational and does not count toward his score. Various methods for controlling these effects are suggested and a hypothetical test format, containing two operational sections and one of two non-operational sections with twelve possible sequences, is presented to illustrate the ways in which context effects may be statistically assessed. Finally, it is recommended that if there are too many sections in a test to adequately assess context effects in one test administration, a data collection plan allowing comparisons over a number of test administrations should be developed. (Author/AEF)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED205595

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
E. E. Stewart
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# METHODOLOGICAL ISSUES RELATED TO THE STUDY OF CONTEXT EFFECTS IN MULTISECTION TESTS*

E. Elizabeth Stewart

Educational Testing Service

The term "within-test context effect" will be used to denote any increment or decrement in performance on a particular kind of test material that is associated with the content of the material administered earlier during the same testing session. Context effects may operate within a separately timed section of a test. A particular question may be easier because it appears in juxtaposition with one question rather than another, or, if the questions in a particular section have a complex format, a question that appears early in the section may be harder than it would be later on. This paper, however, will focus on effects associated with the content of successively presented separately timed sections of a test.

Within-test context effects may be direct, that is, linked to the earlier presentation of the same type of material, or indirect, that is, linked to the earlier presentation of different types of material. Direct context effects are often referred to as practice effects. Context effects can be distinguished conceptually, although not always empirically, from effects due to the sheer amount of prior testing and also from motivation-related effects due to examinees' beliefs as to whether or not a particular test section counts toward their scores.

TM 810 496

*Paper presented at the 1981 Annual Meeting of the National Council on Measurement in Education, Los Angeles, April 14, 1981

There are situations in which a test containing, say five kinds of test material and six separately timed sections, may be administered to one examinee in the order ABCDEe; to a second, in the order BDcEAC; and to a third, in the order aECADB. (Upper- and lower-case representations of the same letter stand for sections containing the same kind of material but different specific questions. Different letters stand for sections containing different kinds of material.) If the three examinees have the same true score on one of the sections, their observed scores could differ because of differences in the conditions under which the section was administered (as well as for other reasons). A practice effect, for example, would be operating if, by virtue of having taken section c prior to taking section C, the second examinee were to earn a higher score on C than the other two examinees.

A canon of the testing profession holds that examinees whose scores will be compared should be tested under the same conditions, so that variations among their scores will not reflect variations among the conditions under which they were tested. Under some circumstances, however, there are compelling reasons for a testing agency to administer a test under conditions that are not uniform for all examinees.

One such circumstance arises from the necessities of collecting data for constructing future editions of the test, for equating scores on different editions, and for conducting special-purpose test-related studies. Frequently, tests used in large-scale testing programs include, for these purposes, both questions that are intended to count toward the examinee's reported score and questions that are not intended to count toward the reported score, sometimes called "nonoperational questions."

In some test administration designs nonoperational questions are intermingled with operational questions in a test or test section administered with a single time limit. In other designs one or more intact sections may consist entirely of nonoperational questions. To serve all the purposes for which nonoperational sections are administered, it is often necessary to administer different kinds of nonoperational questions to different examinees. Otherwise, it would not be possible to collect the variety of data required.

There may also be strong reasons for administering the sections of a test in different sequences to different examinees. In testing programs in which some test sections are nonoperational and in which the operational portions of the test are made available to examinees after the administration, it is important to vary the location of the nonoperational sections. The purpose of varying their locations is to avoid the establishment of a discernible pattern that might enable examinees to identify, while taking a test, the parts that will not contribute to their scores. If examinee performance were affected by the belief that a section "did not count," there could be serious adverse consequences, as, for example, perturbations in the equating of scores on new test editions and a reduction in the test developer's ability to use pretest statistics to control the difficulty of new editions.

The foregoing conditions, then, mean that the kind and amount of test material that precedes a particular test section may be different for different examinees. A conceptual distinction can be made between conditions that are ostensibly different and conditions that are psychologically different. The purpose of studying within-test context effects is to try

to determine the degree of psychological sameness or distinctiveness of conditions that are ostensibly different.

There are two reasons for being concerned about possible within-test context effects. One reason is that they may introduce variance among the scores of individual examinees that is irrelevant to the construct the test is designed to measure. One would like their contribution to the overall standard error of measurement associated with the test scores to be negligible or, at worst, very small. One way to try to minimize within-test practice effects is to make available to all prospective examinees, prior to the test administration, a substantial amount of practice test material and to encourage them to become thoroughly familiar with the instructions and format for each kind of question and to work on the practice material under simulated operational conditions so as to learn to pace themselves appropriately on each separately timed section. This approach is followed in many testing programs.

A second reason for concern about possible within-test context effects is their potential impact on item and test statistics. When the nonoperational test locations are used for pretesting, these effects may be manifested as shifts in the difficulty of test questions from the occasion of their pretesting to their later administration as part of an operational test. When the nonoperational locations are used to collect data for score equating, the effects may appear as systematic displacements of the scaled scores on new test editions.

Effects that contribute to the raw score standard error of measurement for individuals in only a very small way have the potential to distort the calculation and/or the application of score equating results, thereby

introducing an additional component of error variance into the reported scores. The distortion can occur in several ways, depending on the design for collecting equating data and, if a new test edition is pre-equated by any of several methods, on the way the edition is administered when it is equated and when it is used operationally. Finally, under some conditions context effects can accumulate over a series of equatings of successive editions of a test.

In a multisection test in which some sections contain the same kind of material, there are a number of possible sources of context effects, as mentioned earlier. The following may be enumerated: (a) direct context effects, or practice effects, which occur when performance on material of a particular kind is affected by an examinee's having been exposed to the same kind of material earlier in the test; (b) indirect context effects, which occur when performance on material of a particular kind is affected by an examinee's having been exposed to another kind of material earlier in the test; (c) serial position effects, resulting from the amount of material, apart from the kind of material, that has preceded a particular section; and (d) motivational effects, resulting from an examinee's believing that a particular test section will not count toward his or her reported score.

Motivational effects, if they exist and artificially depress performance, can create a problem for examinees who believe incorrectly that a section does not count. Like any other source of score variance not related to the construct being measured, they can also affect the integrity of item and test statistics. They can be controlled by unpredictably varying the location of the nonoperational material in the test, either

for different examinees at the same test administration or from one test administration to another. Maximum capability for administrative control exists when all test sections are the same length.

The other three kinds of effects can be studied either by collecting appropriate data as part of regularly scheduled operational test administrations or by providing for special experimental administrations. For many testing programs, the latter course is feasible only occasionally, if at all, because of difficulties in arranging for the cooperation of sufficient numbers of experimental subjects who are reasonably representative of the examinees who attend operational administrations. Since most effects are likely to be small, large amounts of data are required to isolate them.

An approach to the study of these effects in an operational testing setting can be illustrated by the following example of a hypothetical test that contains two operational sections, A and B, that are taken by all examinees and one nonoperational section, which for any particular examinee will be either a or b. Two of the three test locations can be occupied by operational sections in any one of six ways, the number of permutations of three things taken two at a time. Since the remaining, nonoperational location can be occupied by a nonoperational section in either of two ways, there are 12 possible sequences of three test sections each (AaB, aAB, BAa, BaA, ABa, aBA, ABb, AbB, BAb, bAB, BbA, and bBA). Under some circumstances it might be realistic to plan to administer all 12 sequences to essentially random samples of examinees attending a regular test administration.

If this data collection design were followed and if the operational raw score were the sum of the raw scores on sections A and B, clearly one question of interest would concern the magnitude of the differences among the means on section A + section B for the 12 groups. These differences could be examined through a one-way analysis of variance, including one or more selected contrasts if they appeared indicated on the basis of prior hypotheses or consideration of the results. In the absence of information from previous studies of the hypothetical examination in question, a series of orthogonal contrasts could be identified in advance, as illustrated below.

I. Prior exposure to material similar to A or B vs. no prior exposure to material similar to either

II. Practice on a vs. practice on b

III. Adjacent vs. nonadjacent practice on a when A is last

IV. Adjacent vs. nonadjacent practice on b when B is last

V. Testing in the order AB vs. BA when the operational sections are in the first two locations in the test

VI. Testing in the order AB vs. BA when the operational sections are separated by a section dissimilar to the operational section that it precedes

VII. Taking the operational sections in the first two locations in the test vs. taking them in the first and third locations separated by a section dissimilar to the operational section that it precedes

The groups involved in the contrasts and their associated weights would be as follows.

| | Sequence | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Contrast | AaB | aAB | BAa | BaA | ABa | aBA | ABb | AbB | BAb | bAB | BbA | bBA |
| I | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 |
| II | | 1 | | 1 | | 1 | | -1 | | -1 | | -1 |
| III | | | | 1 | | -1 | | | | | | |
| IV | | | | | | | | 1 | | -1 | | |
| V | | | 1 | | -1 | | -1 | | 1 | | | |
| VI | 1 | | | | | | | | | | -1 | |
| VII | -2 | | 1 | | 1 | | 1 | | 1 | | -2 | |

More detailed analyses at the section score level could be carried out in a variety of ways. One approach would entail separate examinations of direct and indirect context effects. For each kind of test material (A and a or B and b), three independent pairwise assessments of practice effects could be made as shown below for A and a. Note that within each of the three comparisons the testing sequences are identical except for the interchanging of the locations of A and a.

| Comparison | Group | Preceding Material | Exposure 1 | Intervening Material | Exposure 2 |
|---|---|---|---|---|---|
| 1 | 1 | None | A | None | a |
| | 2 | None | a | None | A |
| 2 | 3 | B | A | None | a |
| | 4 | B | a | None | A |
| 3 | 5 | None | A | B | a |
| | 6 | None | a | B | A |

If the investigator were willing to merge the raw scores on A and a, each comparison could be made via a two-factor analysis of variance with

repeated measures on one factor. Willingness to merge the raw scores rests on the assumption that the raw score scale of a can be translated to the raw score scale of A by adding a constant. If the investigator were less willing to make this assumption than to forego the additional precision afforded by the repeated measures design, two *t*-tests of the differences between uncorrelated means could be performed for each comparison, one for A and one for a. The other comparison of interest, which is not orthogonal to those just described, is among scores on A for the three groups that were exposed to a before taking A (and, separately, among scores on a for the three groups that were exposed to A before taking a). These comparisons could be made by one-way analyses of variance.

The effect on A of exposure to B or b could be assessed by a one-way analysis of variance of the scores of groups 7 + 8, 9 + 10, and 11 + 12. Note that the effect of taking A in the first, second, or third location in the test is inseparable from the effect of different amount of exposure to B and/or b prior to taking A.

To facilitate descriptive comparisons across different sets of test materials, all effects should be expressed in units of the appropriate within-group standard deviation.

The preceding example involved a three-section test and thus represented a situation in which it very well might be feasible to collect data on all possible sequences of test sections at a single administration. The feasibility of simultaneously collecting all data of interest diminishes as the number of sections in the test increases. Consider, for example, a five-section test in which three sections, containing test material A, B, and C, are operational and two sections are nonoperational. All examinees

take the same operational test material. Any particular examinee takes two of the three nonoperational counterparts (a, b, and c) of the operational sections. In this situation the number of possible sequences of five sections is quite large. Three of the five test locations can be occupied by operational sections in any one of 5!/(5 - 3)!, or 60, different ways. After the operational sections have been positioned, the two locations left for nonoperational sections can be filled in any one of 3!/(3 - 2)!, or 6, different ways. Thus, the three operational and two nonoperational sections could be administered in any one of 360 sequences. The number of sequences is even greater in the case of six- to eight-section tests, which are used in some testing programs.

Because the number of examinees available to take any particular sequence would be undesirably small or because the practical complexities of preparing, administering, and processing so many different versions of the test could not be dealt with, the possible sequences may be too numerous to be given at a single test administration. The study of context effects then requires the development of a data collection plan that will allow different subsets of comparisons to be made at different administrations, based in many cases on different but parallel collections of test questions.

Generalized guidelines for the development of a data collection plan cannot be given because the situational constraints tend to be test-specific. The recommended approach is to list the effects hypothesized to be potentially the most serious and to determine how they might be examined over a series of test administrations, given the situational constraints.

Descriptive comparisons across administrations could be made when all effects had been expressed in standard deviation units.